Clerk of the Superior Court

Information Technology Group (ITG)

Statewide eFiling

Standard XML File Format Adoption White Paper

Version: 1.00 Revision Date: 11/30/2009

Table of Contents

REVISION HISTORY

Version	Comments	Author	Date
1.0	Initial Release	Ron Bitterli Mark Jensen	11/30/2009

1. Introduction

On November 17, 2009 the members of TAC were asked to provide comments relating to the OpenDocument Format. The email contained the following two statements.

> 1) *COT has sent the XML format decision back to TAC accompanied by the business drivers members agreed upon and insisted on a formal motion. Enough reservation existed in October that no motion was made. Karl has asked that you forward your issues about OpenDoc by November 25 to allow staff time to compile a list of items to address ahead of the meeting. **COT's stated business drivers are cost and space required to store and transmit electronic documents, as well as searchability.***

> 2) *In case one of your issues is the relative size of the pdf compared to odt presented in the white paper, I have attached the Word documents from which the pdfs were built so you can now engineer your own smaller sized pdfs and describe in detail the settings used to optimize the file sizes. Please return your pdfs to me along with the detailed description of exactly how they were created. Again, getting these by November 25 will allow for a comparison table to be created for the meeting. My converted file sizes for basic pdf and searchable pdf were:*
> _E-court (3 pages) 46 KB and 118 KB
> _Test (13 pages) 103 KB and 516 KB
> _Long (100 pages) 559 KB and 4390 KB

This document provides the Maricopa County Clerk of the Superior Court response to the question of adopting a standard XML file format document type and provides discussion information addressing the intent to use this new format in the statewide eFiling initiatives currently under construction.

The AOC is currently proposing the adoption of the OpenDocument Format (ODF) for the XML standard currently identified as XML in ACJA § 1-506.

The OpenDocument Format (ODF) is an XML-based file format for representing electronic documents such as spreadsheets, charts, presentations and word processing documents. The .odt file extension represents word processing (text) documents defined by the OpenDocument Format (ODF). For consistency, the use of the term "ODF" in this document refers to word processing (text) documents with the .odt extension.

There are two distinct topics that have been presented to the TAC for consideration.

1) the topic of adopting a standard XML file format

2) the topic of adopting that file format as the standard for the exchange and storage of documents within the eFiling initiatives

The following sections address each of these topics separately.

2. Adopting a standard XML file format

From the minutes of the November 6th COT meeting:

> *"Justice Hurwitz set the stage for Karl Heckart's presentation by defining the question as not whether XML should be used for electronic documents but rather what brand of XML should be the courts' standard to fill in the current hole in ACJA § 1-506….Justice Hurwitz clarified that AJC long ago decided there would be XML submissions, but left the details for later. Rather than debate the wisdom of XML versus other formats, TAC now needs to fill in the details."*

It would be very helpful to continue to frame the discussion in this manner. To this end, the TAC should submit to the COT for approval a proposed revision of ACJA 1-506 which amends the section on Document Specifications and specifies the type of XML they are recommending be adopted as a document standard along with PDF.

Currently the code reads:

> *"Documents shall be in a format that provides for browser accessibility and no material alteration to content or appearance. Documents shall be formatted in either:*

- *PDF (Portable Document Format) version 2.x or higher; or*
- *XML (Extensible Markup Language), after the supreme court adopts standards for its use."*

Research is still needed in order to determine whether there is a conclusive "best" XML format and whether the technology is ready to be adopted as a standard.

3. Exchange and Storage of Documents

In the eFiling business function, the process is more about document exchange and storage rather than document editing. The issue can be framed around appropriateness for the particular business function. **ODF** is an OPEN standard format for document **editing** of text documents (ISO/IEC 26300:2006), **PDF** is the current OPEN standard format for **document exchange** (ISO 15930-1:2001, ISO 19005-1:2005, ISO 32000-1:2008).

3.1 Standard Format for Word Processors

If we were going to standardize on a Word Processor to use for editing documents for the Legal Industry, we would certainly consider ODF but, we would most likely not choose ODF as the standard that we employ due to the low market share that it currently holds. We would more likely consider MS Word (doc) due to its widespread use. Over the next few years, ODF may become more widely used which would make it more feasible for adoption for document editing. But, since we are talking about storage and distribution of court documents, PDF is the current standard and there is no foreseeable change to that standard.

3.2 Standard Format for Document Exchange and Storage

Clarifying the XML document standard is one thing, but prescribing its adoption and the exclusion of other viable formats is another thing entirely. PDF is the de facto standard in all current eFiling implementations in Arizona, and across the country. Forcing the adoption of a new and unproven file format by electronic filers in Arizona is not consistent with good customer service, and will most likely create animosity and resistance to the adoption of eFiling.

Adoption of ODF may occur over time if it is truly a superior format that affords real business advantages for the court and its eFilers. If it is approved as an optional document standard, its utility can be established and studied in the context of an actual eFiling pilot implementation.

3.3 Preserving Document Content and Appearance

The designation of PDF as one of the document standards endorsed in ACJA 1-506 did not come about through any misconception as to its utility for affording protection against tampering or manipulation. Document security and authentication are addressed in section 1-506 (E). The advantages of PDF pertain to its ability to preserve the content and appearance of an original source document when printed or viewed in a browser, its ubiquity in electronic filing implementations throughout the nation, and its ease of access and presentation via software and readers which are free and readily available.

ODF files are not guaranteed to print or have the same appearance across all word processors as certain fonts and special features (bullet styles, formatting styles, etc.) may not be supported.

Although security of the document content cannot be guaranteed by the selected format, unintentional changes can be made to ODF files during the viewing process, by deleting or adding text accidentally. With the PDF format, there would have to be a deliberate malicious attempt to modify the content of the document.

3.4 Document Editing

One of the arguments for selection of ODF as the standard has been that it will allow eFiled documents to be edited and provide JO's the ability to modify the content of the documents as needed. This gets into all of the very detailed and complex requirements around what JO's really need in terms of editing and searching documents. The majority of eFilings, aside from proposed orders, should not be edited. They are already accepted eFilings, in the Electronic Court Record (ECR), and should not be altered, even by the bench.

At this time, there does not seem to be a compelling need to use ODF when the majority, if not all, of the JO's utilize Word. Allowing proposed orders to be filed as Word documents along with court documents in PDF format would provide the ability for the JO's to create the final orders from the proposed orders and preserve them as PDF files in the ECR.

4. File Size, Storage and Bandwidth

The issues of file size, storage and bandwidth have been raised in the context that "smaller is better" since it will lead to less storage requirements and less impact to bandwidth. Although these are certainly considerations they are not necessarily driving factors in deciding what standard to implement.

4.1 File Size Issue

Comparison of relative file sizes could certainly be a consideration if the business case and processing scenarios have been fully laid out to show how and where the ODF file type will be used, where and when it will be transmitted, and at what points conversion to another format may be required.

In addition, the larger file sizes being ascribed to PDF has not been shown to negatively impact current efiling and EDMS infrastructures.

No negative impact has been observed to date in any implementation anywhere that has caused a court to reconsider PDF as its document standard, and no analysis of historic data has been provided to quantify or substantiate these concerns. (An example of such an assessment performed on behalf of the Clerk of Superior Court in Maricopa County for its EDMS Master Plan is available, which, if modified and applied to the statewide environment, may or may not help to substantiate existing concerns.)

As to the file sizes of ODF files compared to PDF files, ODF files are indeed smaller than their comparable PDF file. The difference does depend upon the conversion process used. Maricopa County has been working with PDF files for the past 7 years and has not found the difference in file size to cause us any concern about storage capacity. The samples offered in the November 17 email were very limited examples and may not reflect the typical documents being eFiled. No analysis was provided as to how these documents would scale to the statewide eFiling project.

4.2 Bandwidth

As to the bandwidth issue, certainly larger files require higher bandwidths, but again, with the differences in current file sizes, the bandwidth requirements would be minimal. Billions of large documents and video files are sent across the internet daily that do not cause bandwidth issues. If the file size differences

that are being discussed are truly expected to cause bandwidth issues, it would be a better conversation to discuss improvements to the current network infrastructure. No analysis has been provided that would demonstrate that bandwidth would become an issue based on the eFiling of PDF documents compared to ODF documents.

4.3 Availability

Although MS Word allows saving as an ODF file, this is only available with MS Word 2007 and does not fully support all of the functionality that MS Word provides, therefore, firms with template documents may have to change their forms to allow them to be saved correctly in an ODF format.

On the other hand, since PDF has been around for 15 years, is widely adopted, and is used by all Maricopa County filers, there would be no difficulties in continuing accepting PDF files. Since PDF is also widely adopted across the nation, there should be no difficulties in retaining PDF as the standard.

5. Summary

There has been no compelling evidence presented to date that would mandate the switch to the OpenDocument Format (ODF) as a standard for document exchange and storage. The widespread use of PDF and Word document formats far outweigh the minimal use of the OpenDocument Format. If a large shift in the next few years can be demonstrated, it may be a topic to revisit at that time.

Shifting the legal community to use ODF editors and deal with ODT files when there are a relatively small number of users familiar with ODF would require everyone in the legal community to make a change. There is not enough compelling evidence to support creating a new standard with such little adoption.

Currently, TurboCourt uses PDF as their national standard. It has been stated that "although TurboCourt uses PDF, Arizona will be standardizing on ODF". I strongly recommend against becoming the leader in changing the current standard. If and when the industry begins to take this direction and it has implementations that can be reviewed and processes with best practices in place, it would then be time to consider changing the current standard.

At this time even a recommendation of ODF as the standard XML format would be difficult to make as no definitive information has been provided detailing why ODF would be the right choice.   Additional work needs to be done to provide TAC with the ability to make a recommendation as to the standard XML format.   As to adopting ODF as an acceptable eFile document format, research and planning will be needed to determine whether this is a viable document format for court records.

6. Supporting Information

The following supporting information is provided for your additional information.

6.1 ISO Press Release

The ISO has issued a press release about the new standard (named "ISO 32000-1:2008"), along with a quote from Adobe Chief Technology Officer Kevin Lynch about the move expanding the PDF universe. "As governments and organizations increasingly request open formats, maintenance of the PDF specification by an external and participatory organization will help continue to drive innovation and expand the rich PDF ecosystem that has evolved over the past 15 years," Lynch said. It's nearly verbatim with what he said back in the AIIM hand-off, but holds true to what typically happens when any file format is ISO certified. They'll typically become more attractive to governments and large corporate customers.

6.2 From Adobe's Website

PDF has been adopted by the International Organization for Standardization (ISO) as the standard for archiving electronic documents. One by one, more than a half billion people who use their computers and mobile phones to view, print, and collaborate on documents and forms have come to trust PDF.

6.3 Excerpt from the AIIM Standards document

The following is a section from the AIIM Standards document relating to TIFF documents:

> ***AIIM ARP1-2009 – Analysis, Selection, & Implementation of Electronic Document Management Systems***
>
> ***(EDMS)***
>
> ***5.4.1.4 Image formats***
>
> *The organization should ensure that all information being scanned, or electronically received is stored in an industry accepted format such as JPEG, JBIG, JPEG 2000, or PDF-A. Non-standard or proprietary file formats should not be used. Caution should be exercised if using TIFF. While TIFF is commonly used, there are multiple problems associated with the ability of the application to use non-standard headers, or tags that 1) may not be documented and/or 2) the misuse of other basic headers, or tags. Additionally, TIFF images can be modified without user knowledge though numerous freely available editing tools. Image formats such as PDF-A are non-modifiable through the file format structure along with the use of "checksums" that should be stored in the document management system as an additional method of ensuring that the file has not been altered, modified, or deleted during the information lifecycle. Non-standard or proprietary formats include any formats used by a single vendor/source and not accepted as a standard file format at either a national or international standards level. Proprietary file formats include but is not limited to:*
>
> - *File formatting that utilizes "file-wrappers" to encapsulate standard file formats within a nonstandard structure,*
> - *TIFF formats that are not fully documented by the vendor and independently verified by the organization to ensure proprietary information is not contained in any of the headers,*
> - *Non-standard file formats not used by multiple vendors/integrators, etc.*
>
> *It is important to note that the industry has found that using PDF-A as the output format for any hardcopy conversion to electronic format eliminates many of the commonly seen problems found with TIFF formatted information including: prevents alteration, incorporates the concept of checksums, all information is fully contained, and the PDF-A format is fully standardized and supported by almost every EDMS solution provider, including all the major document imaging solutions currently available (with the exception of smaller solutions that still rely on proprietary methods and concepts).*